Streamlining M&V Through Automated Analytics

First Meeting of the Technical Advisory Group
June 12, 2014

Funded by US Department of Energy Building Technologies Office, Cody Taylor

Research Team:
Jessica Granderson, Michael Sohn, Erin Hult, , Phillip Price, Jodi Bellacicco, LBNL
David Jump, QuEST

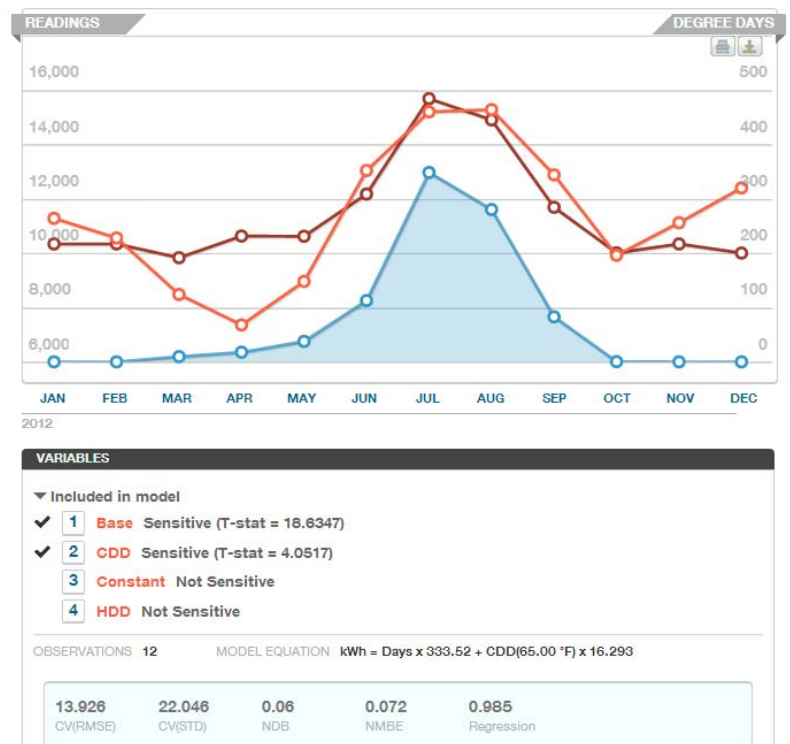
BERKELEY LAB

Meeting Agenda

- Introductions
- Project Motivation, Scope, and Objectives
- Planned Engagement with the TAG
- Testing Procedure for Baseline Model Accuracy
- Metrics to Quantify Baseline Model Performance
- Time Horizons for Model Training and Prediction

Motivation

- **High level goal**: Enable the industry to harness emerging tools and devices to conduct M&V at dramatically lower cost, with comparable or improved accuracy



Who is the Audience for This Work?

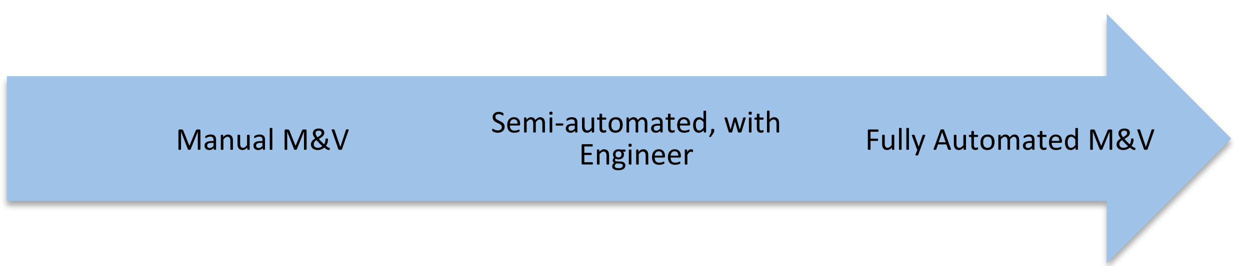
Organization	Event	Date of Event
CEE Whole Buildings Committee	Industry partners meeting, Winter meeting day-ahead workshop, ongoing committee meetings	September 2013, January 2014, ongoing
SEE Action EM&V Working Group	Working group meetings	December 2013, May 2014
CA PUC	EM&V Quarterly Meeting	March 2014, September 2014
ACEEE	Market Transformation, Summer Study paper and informal session	March 2014, August 2014
ESource	Emerging Technologies Leadership Group, Annual Forum,	April 2014, September 2014
NEEA, NEEP, MEEA	Webinar	May 2014
Analytical tool vendors, developers	Webinar	June 2014
AESP	Brown Bag seminar, Summer Conference, National conference	August2014, February 2015
ASHRAE	Summer Meeting panel session	July 2014
Greenbuild	M&V Panel Session	October 2014 (wait listed)
IEPEC	Annual Conference	September 2015

What Questions Are Being Asked?

- Can I use a whole-building approach for my programs and projects?
- How can I determine whether a given model or commercial tool is robust and accurate?
  - In general, and for a specific portfolio of buildings?
- What repeatable test procedures can be used to evaluate model and tool performance, and which metrics provide critical performance insights?
- How can I compare and contrast proprietary tools and ‘open’ modeling methods for M&V?
- How can we reduce the time and costs necessary to quantify gross savings?

Scope of Current Project Activities

- Whole-building and system-level assessment of avoided energy use
  - With implications for normalized savings approaches
- Streamlining and scaling M&V in practice:
  - Analysis of fully automated baseline model capabilities
  - Establishes a *floor* of performance that can be improved by the oversight of engineer, used to reduce costs and time



Project Objectives

- Raise industry awareness through extensive outreach at national level
- Refine testing procedure established in past work
- Solicit novel baseline models from the public
- Apply test procedure to evaluate those models
- Publish and disseminate results to support increased adoption of M&V

Planned Engagement with the TAG

1. Kickoff, June 2014
   – testing procedure, metrics, time horizons

2. September 2014
   – models obtained in the solicitation, and test data sets acquired by the team
   – time horizons (cont.)
   – format to present results, e.g. plots, tables, 'scorecards', etc

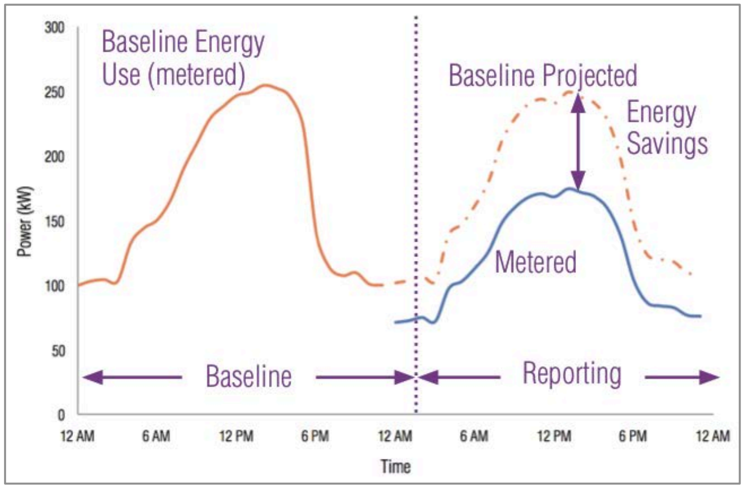
3. Dec/Jan 2014 preliminary model testing results

4. Final findings and next steps webinar, open to audiences beyond the TAG

Testing Procedure for Baseline Model Accuracy

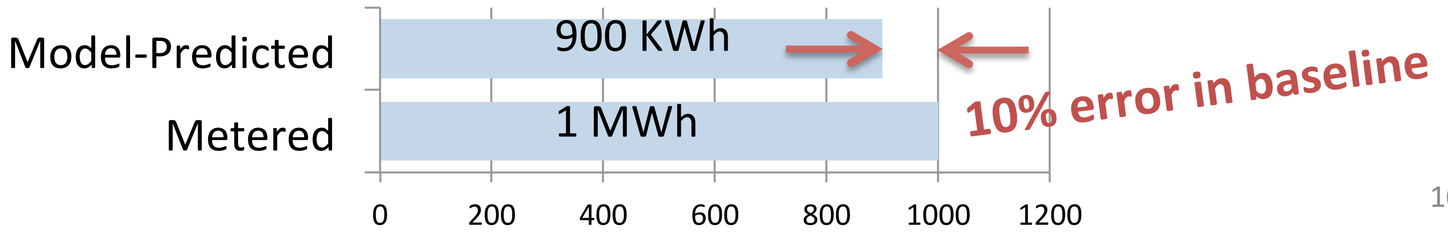
How Accurate Is the Baseline Model?

M&V Use Case



Error in reported savings is proportional to error in baseline projection

Error = **% difference** between total **metered** energy use, total **model-predicted use**



Testing Procedure

Baseline Model

Test Data:
Many buildings, metered data

Model

- **Split data set** into training & prediction period.
- **Train the model** by showing data, hiding prediction-period data;
- **Generate post-period predictions**

Compare

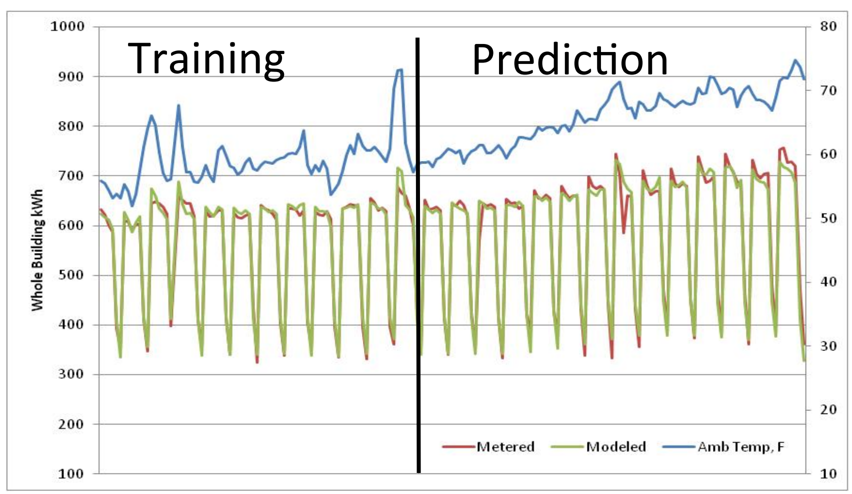
Compare predicted data to actual data that was 'hidden' from model to quantify error.

Repeat for many buildings

Assess

Calculate Performance Metrics, e.g. %Error, R^2, CV(RMSE) …

Outdoor Temp
Metered
Modeled



Testing Procedure

- Statistical cross-validation based on large data sets comprising many buildings, with a ~18-month history of consumption data
- Approach takes advantage of increasing availability of building energy data
- Approach provides more insight into predictive accuracy than using a static pre-period for a single building
  - Exposes models to more wider range of building behavior
  - Provides population-level insights for predictability, screening
- Focus is on model predictive accuracy for target buildings or portfolios of interest; this accuracy is proportional to the error in total calculated savings once a project is implemented.
- Parallel complementary work (PGE-funded) addresses how to most robustly quantify *uncertainty in achieved savings* for a given building, once a model has been selected for use, and an improvement has been made

Example of Results Generated from Applying the Testing Procedure

- Distributions of error and model fitness metrics

- Tabular example showing percentiles and mean of absolute percent bias error
  - 5 models, whole-building electric data, 12-month training period, 12-month prediction period

- Plots, other presentations are possible (TAG meeting #2)

Model	10%	25%	50%	75%	90%	Mean
Mean Week	0.82	2.21	4.82	9.63	19.42	8.40
Monthly CDD and HDD	0.69	2.09	4.53	10.03	19.38	8.46
Day, Time, and Temperature	0.69	2.17	4.51	9.26	19.41	8.42
Day and Change Point	0.73	2.02	4.70	9.22	18.84	8.24
Time of Week and Temperature	0.82	2.21	4.82	9.63	19.42	8.40

Questions on the Testing Procedure

- Questions for the LBNL/QuEST Team?
- Questions for the TAG:
  - Are there any critical technical flaws to this approach to characterize baseline model performance?

* “This approach” = statistical cross-validation based on large data sets with a year+ of historic consumption and independent variable data

Questions on the Testing Procedure

- Questions for the TAG (cont.):
  - If you saw critical technical flaws, what are they, and how might they be remedied?

Metrics to Quantify Baseline Model Performance

Approach To Selecting Performance Metrics

- Many metrics are used and discussed in the literature
- Some are quite similar, e.g., total bias (directional/signed) and total error (absolute value)
- Others provide distinctly different insights into predictive performance e.g., total bias and RMSE
- Prior work showed that focusing on a couple of *primary* metrics aids in tractability and interpretation of results,
  - Particularly for less technical audiences, who are often users of the results
  - Secondary metrics can also be calculated, and presented in appendices for audiences who desire more detail

Metrics Discussion Paper

- Overview of metrics from literature, discussion of different insights provided, relevance to M&V use case

- English language and equation-based definitions of
  - Total bias, total error, mean bias
  - Total normalized bias
  - Mean absolute percent error
  - Normalized mean bias error
  - Root mean squared error, CV(RMSE)
  - Coefficient of determination (R-squared)
  - t and F statistics

Mathematical Definition of Common Metrics

$$Total\ Bias = \sum_{i}^{N}(y_i - \hat{y}_i)$$

$$Total\ Error \sum_{i}^{N}|y_i - \hat{y}_i|$$

$$Mean\ Bias = \frac{1}{N}\sum_{i}^{N}(y_i - \hat{y}_i)$$

$$Total\ Normalized\ Bias = \sum_{i}^{N}\frac{(y_i - \hat{y}_i)}{y_i} \times 100$$

$$Mean\ Absolute\ PercentError = \frac{1}{N}\sum_{i}^{N}\frac{|y_i - \hat{y}_i|}{y_i} \times 100$$

$$Normalized\ Mean\ Bias\ Error = \frac{\sum_{i}^{N}\frac{(y_i - \hat{y}_i)}{y_i}}{\bar{y}} \times 100$$

$$Root\ Mean\ Squared\ Error = \sqrt{\frac{1}{N}\sum_{i}^{N}(y_i - \hat{y}_i)^2}$$

$$CV\ Root\ Mean\ Squared\ Error = \frac{\sqrt{\frac{1}{N}\sum_{i}^{N}(y_i - \hat{y}_i)^2}}{\bar{y}} \times 100$$

normalized RMSE, w large n

$$Coeff.of\ Determ., R^2 = 1 - \frac{\frac{1}{N}\sum_{i}^{N}(y_i - \hat{y}_i)^2}{\frac{1}{N}\sum_{i}^{N}(y_i - \bar{y}_i)^2}$$

Candidate Metrics of Focus

1. Total normalized bias

$$Total\ Normalized\ Bias = \sum_{i}^{N} \frac{\left(y_i - \hat{y}_i\right)}{y_i} \times 100$$

- Percent difference between total model-predicted energy use and total actual energy use
- Clear relevance to errors in reported savings
- Normalization aids in simultaneous treatment of both large and small building loads
- Bias retains directionality of differences, i.e., under or over-prediction, which has implications for savings payouts and incentives

Candidate Metrics of Focus

2. Coefficient of variation of the root mean squared error

$$CV\ Root\ Mean\ Squared\ Error = \frac{\sqrt{\frac{1}{N}\sum_{i}^{N}\left(y_i - \hat{y}_i\right)^2}}{\bar{y}} \times 100$$

**normalized RMSE, w large n*

- Squares difference between predictions and data to highlight large differences between predictions and data
- Favors models that predict the overall shape of the energy meter time series
  - Added insight for extrapolation as in normalized savings calculations
- Prominent in industry references such as Guideline 14

Metrics Questions for the TAG

- Do you agree that focusing on a primary set of just a few metrics is a viable approach?
- Are there important aspects of performance that we would miss with a focus on total normalized bias and CV(RMSE)?

Metrics Questions for the TAG (Cont.)

- If you thought there are important aspects of performance that we would miss with a focus on total normalized bias and CV(RMSE), which additional metrics would you add, and why?

Time Horizons

Approach To Selecting Time Horizons

- Testing procedure requires definition of time duration of pre- and post- periods over which models are trained, and predictions are generated

- As with metrics, any number of time horizons can be selected for focus
  - we would like to focus on those most critical to whole-building and system-level savings approaches

- To do so, we consider time frames used in typical practice, as well as references such as Guideline 14

Time Horizons for Whole-Building Savings

- Current guidance for whole-building approaches is to use 12 months pre- (training) and post- (prediction) data

- Prior work showed that with a fixed 12-month prediction period
  - Monthly and interval models performed ~equally with 12 months training data
  - Performance for interval models ~maintained with 6 months training data, performance of monthly models degraded
  - No model performed well with only 3 months training data, but differences in relative model performance clearly emerged

Considerations in Selecting Time Horizons for Whole-Building Savings

- Conceptually, ‘break points’ as in previous example provide compelling, understandable insights

- Keeping either the prediction or training period fixed aids in interpretation of results
  - Difficult to draw conclusions in comparing, e.g., 12/6 with 6/3 with 3/12

- We know there is desire to shorten total time required for M&V
  - Is a 12 month prediction prediction period critical, based on current practice?

Question for the TAG

- What pre- and post- time horizons do you recommend for testing the accuracy of whole-building savings approaches?
  - Consider ~3 cases of highest interest

Time Horizons for System-level Savings

- Time horizons for system level assessments are often shorter in practice, than those for whole-building
  - Weeks to months

- Heating and cooling loads are dependent on season, with implications for *which* months are used in addition to *how many* months are used

- We are considering 3 to 6 months

Question for the TAG

- What pre- and post- time horizons do you recommend for testing the accuracy of system-level savings approaches?
  - Consider ~3 cases of highest interest for heating and cooling loads

Next Steps

- We will circulate a summary of takeaways from this meeting
- July release of Solicitation of novel models, August selection
- Next TAG meeting in September
  - models obtained in solicitation, test data sets acquired
  - time horizons (cont.)
  - format to present results, e.g. plots, tables, 'scorecards', etc

Thank You For Participating!